PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, C12N 1/20, 15/00, C07H 21/04	A1	(11) International Publication Number: WO 97/12993
		(43) International Publication Date: 10 April 1997 (10.04.97)

(21) International Application Number: PCT/US96/14655

(22) International Filing Date: 12 September 1996 (12.09.96)

(30) Priority Data:
60/004,664 2 October 1995 (02.10.95) US

(71) Applicant: THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY [US/US]; Suite 350, 900 Welch Road, Palo Alto, CA 94304 (US).

(72) Inventors: COX, David, R.; 2743 Hallmark Drive, Belmont, CA 94002 (US). FAHAM, Malek; Apartment #17, 275 Ventura Avenue, Palo Alto, CA 94306 (US).

(74) Agents: BOZICEVIC, Karl; Fish & Richardson P.C., Suite 100, 2200 Sand Hill Road, Menlo Park, CA 94025 (US) et al.

(81) Designated States: AU, CA, JP, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: MISMATCH REPAIR DETECTION

(57) Abstract

Mismatch Repair Detection (MRD), a novel method for DNA-variation detection, utilizes bacteria to detect mismatches by a change in expression of a marker gene. DNA fragments to be screened for variation are cloned into two MRD plasmids, and bacteria are transformed with heteroduplexes of these constructs. Resulting colonies express the marker gene in the absence of a mismatch, and lack expression in the presence of a mismatch. MRD is capable of detecting a single mismatch within 10 kb of DNA. In addition, MRD has the potential for analyzing many fragments simultaneously, offering a powerful method for high-troughput genotyping and mutation detection in a large genomic region.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	**GB**	United Kingdom	**MW**	Malawi
AT	Austria	**GE**	Georgia	**MX**	Mexico
AU	Australia	**GN**	Guinea	**NE**	Niger
BB	Barbados	**GR**	Greece	**NL**	Netherlands
BE	Belgium	**HU**	Hungary	**NO**	Norway
BF	Burkina Faso	**IE**	Ireland	**NZ**	New Zealand
BG	Bulgaria	**IT**	Italy	**PL**	Poland
BJ	Benin	**JP**	Japan	**PT**	Portugal
BR	Brazil	**KE**	Kenya	**RO**	Romania
BY	Belarus	**KG**	Kyrgystan	**RU**	Russian Federation
CA	Canada	**KP**	Democratic People's Republic of Korea	**SD**	Sudan
CF	Central African Republic			**SE**	Sweden
CG	Congo	**KR**	Republic of Korea	**SG**	Singapore
CH	Switzerland	**KZ**	Kazakhstan	**SI**	Slovenia
CI	Côte d'Ivoire	**LI**	Liechtenstein	**SK**	Slovakia
CM	Cameroon	**LK**	Sri Lanka	**SN**	Senegal
CN	China	**LR**	Liberia	**SZ**	Swaziland
CS	Czechoslovakia	**LT**	Lithuania	**TD**	Chad
CZ	Czech Republic	**LU**	Luxembourg	**TG**	Togo
DE	Germany	**LV**	Latvia	**TJ**	Tajikistan
DK	Denmark	**MC**	Monaco	**TT**	Trinidad and Tobago
EE	Estonia	**MD**	Republic of Moldova	**UA**	Ukraine
ES	Spain	**MG**	Madagascar	**UG**	Uganda
FI	Finland	**ML**	Mali	**US**	United States of America
FR	France	**MN**	Mongolia	**UZ**	Uzbekistan
GA	Gabon	**MR**	Mauritania	**VN**	Viet Nam

MISMATCH REPAIR DETECTION

GOVERNMENT GRANTS

This invention was made with government support under Contract Nos. HD 24610 07-10 and 5T32GM07618 awarded by the National Institutes of Health. The Government has certain rights in this invention.

INTRODUCTION

Technical Field

The field of this invention is genetic mapping.

Background

The detection of mutations in genomic DNA plays a critical role in efforts to elucidate the genetic basis of human disease. For many types of genetic screening and analysis, knowledge of the presence of a mutated copy of a gene is essential. Such information may be used in prenatal and other genetic testing, as well as analysis of tumor cells and other somatic mutations. For many genes, there are a number of different mutations that can affect function. While many approaches are currently applied to the problem of mutation detection, no single technique provides a rapid method for screening large stretches of genomic DNA with high sensitivity and specificity.

Current techniques for detecting unknown mutations in genomic DNA fall into three general classes. The first class of techniques, which includes single strand conformational polymorphism (SSCP) analysis, denaturing gradient gel electrophoresis (DGGE), and heteroduplex analysis in gel matrices, detects conformational changes created by DNA sequence variation as alterations in electrophoretic mobility. These techniques are limited by the need to determine optimum reaction conditions for each DNA fragment and by a marked decrease in sensitivity with increasing DNA fragment size. The

second class of techniques, which includes RNaseA cleavage, chemical mismatch cleavage (CMC) and enzyme mismatch cleavage (EMC) uses chemicals or proteins to detect sites of sequence mismatch in heteroduplex DNA. These techniques can be used to assay many different DNA fragments with a single set of assay conditions. In addition, these techniques can be used to detect mutations in larger DNA fragments. However, even with this second class of techniques, the upper limit for the size of the screened DNA fragment is about 1 kb.

A method termed "Genomic Mismatch Scanning" (GMS) has been used to identify regions of the genome identical by descent (U.S. Patent no. 5,376,526). However, GMS yields a probe only for regions of identity. The ability to utilize probes for both regions of identity and regions of difference allows for an improved signal to noise as compared to the use of a single probe.

In view of the importance of genetic testing, methods whereby one can easily screen for genetic mismatches between two DNA molecules is of great interest. A preferred method could provide for isolated DNA samples of regions of identity and regions of difference. A simple method to determine whether two DNA molecules are identical or different would also be advantageous.

Relevant Literature

Techniques for detection of conformational changes created by DNA sequence variation as alterations in electrophoretic mobility are described in Orita *et al.* (1989) P.N.A.S. **86**:2766; Orita *et al.* (1989) Genomics **5**:874; Myers *et al.* (1985) N.A.R. **13**:3131 (1985); Sheffield *et al.* P.N.A.S. **86**:231; Myers *et al.* Meth. Enzym **155**:501; Perry and Carrell (1992) Clin. Pathol. **45**:158; White *et al.* (1992) Genomics **5**:301.

Techniques that use chemicals or proteins to detect sites of sequence mismatch in heteroduplex DNA are described in Cotton *et al.* (1988) P.N.A.S. **85**:4397; Myers *et al.* (1985) Science **230**:1242; Marshal *et al.* (1995) Nature Genetics **9**:177 (1995); Youil *et al.* (1995) P.N.A.S. **92**:87.

Grompe (1993) Nature Genetics **5**:111 reviews methods for screening large stretches of DNA. Mapping strategies may be found in Risch (1990) Am. J. Hum. Genet. **46**:229-241; Lander and Botstein (1987) Science **236**:1567-1570; and Bishop and Williamson (1990) Am. J. Hum. Genet. **46**:254-265. Sandra and Ford, (1986) Nucleic Acids Res. **14**:7265-7282 and

Casna, *et al.* (1986) Nucleic Acids Res. **14**:7285-7303 describe genomic analysis.

However, several approaches are presently available to isolate large DNA fragments, including long range PCR with enzymes with high fidelity described in Nielson *et al.* (1995) Strategies **8**:26; recA-assisted cleavage described by Ferrin and Camerini-Otero (1991) Science **254**:1494; and the use of a single set of oligonucleotide primers to PCR amplify multiple specific fragments simultaneously in Brookes *et al.* (1995) Human Molecular Genetics **3**:2011.

The E. coli methyl mismatch repair system is described in Wagner and Messelson (1976) P.N.A.S. **73**:4135; Modrich (1991) Annu. Rev. Genet. **25**:229; Parker and Marinus (1992) P.N.A.S. **89**:1730; and Carraway and Marinus (1993) J. Bacteriology **175**:3972. The normal function of the E. coli methyl-directed mismatch repair system is to correct errors in newly synthesized DNA resulting from imperfect DNA replication. The system distinguishes unreplicated from newly replicated DNA by taking advantage of the fact that methylation of adenine in the sequence GATC ocurs in unreplicated DNA but not in newly synthesized DNA. Mismatch repair is initiated by the action of three proteins, MutS, MutL and MutH, which lead to nicking of the unmethylated, newly replicated strand at a hemimethylated GATC site. The unmethylated DNA strand is then digested and resynthesized using the methlated strand as a template. The methyl-directed mismatch repair system can repair single base mismatches and mismatches or loops of up to four nucleotides in length. Loops of five nucleotides and larger are not repaired.

SUMMARY OF THE INVENTION

Compositions and methods are provided for an *in vivo* bacterial assay, termed "Mismatch Repair Detection" (MRD). The method detects mismatches in a double stranded DNA molecule, where the sequence of one strand differs from the sequence of the other strand by as little as a single nucleotide. The two strands of the DNA molecule are from different sources. One strand is unmethylated DNA, having a detectable marker gene and the sequence being tested for mismatches. The other strand is methylated DNA, having an inactivated copy of the marker gene where the defect does not activate repair mechanisms, and another copy of the sequence of interest. Heteroduplex dsDNA formed from the hybridization of the two strands is transformed into a bacterial host with an active methyl mismatch repair system (MMR host).

The host repair system is activated by a mismatch in the sequence of interest, and will then "co-repair" the marker gene, to produce an inactive, double stranded copy. When the two strands of the sequence of interest are a perfect match, the marker gene is not altered, and the transformed bacteria will produce active marker. Where a mismatch is present, the transformants are readily identified by the lack of active marker, and may then be isolated and grown for further analysis. MRD is a rapid method for analysis of numerous fragments simultaneously. It is useful as an assay for enumerating differences between various sources of DNA, and as a means of isolating DNA with variant sequences.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 depicts the method for mismatch repair detection.

Figure 2 depicts the method using single or double stranded vectors and an amplification product as a test sequence.

Figure 3 shows a plasmid map of pMF200 and pMF100.

Figure 4 depicts formation of heteroduplex DNA

Figure 5 depicts analysis of MRD results by hybridization.

DESCRIPTION OF THE SPECIFIC EMBODIMENTS

Mismatch Repair Detection (MRD) is a method of detecting mismatches in the sequence of a double stranded DNA molecule. The method will determine whether two DNA sequences differ by as little as a single base change, in a region of over 10,000 nucleotides. Multiple DNA fragments can be analyzed in a single reaction, and the process is easily scaled up to run large numbers of reactions in parallel. Depending on the input DNA, MRD can be used for various purposes. It is used in genetic mapping to analyze large regions of eukaryotic chromosomes for the presence of mutations. In a large pool of genomic or cDNA clones, the method will identify those DNAs where there is a mismatch between the control and test population, providing a particularly simple method of isolating variant alleles from a particular locus or region. The method can also be used to detect somatic changes in DNA, such as those found in tumor cells, or in the hypermutation of antibody genes. A key advantage of MRD is that, once provided with suitable vectors, the procedure is extremely easy to perform.

MRD Method

MRD exploits the ability of bacterial cells to "co-repair" long stretches of DNA. When the two strands of a dsDNA molecule have a mismatch, *i.e.* the nucleotides at a specific position are not complementary, the methyl-directed mismatch repair system of a bacteria will excise and replace the incorrect nucleotide. The strand of DNA that contains within it the modified sequence motif GA^{methyl}TC is recognized by the repair system as the "correct" sequence. Correction is initiated by mismatches of one to four contiguous nucleotides. A loop of 5 or more mismatched nucleotides is not recognized by the proteins responsible for initiation of repair, and will remain uncorrected in the absence of other mismatches. However, if repair is initiated at one site on the DNA molecule, then a region extending for at least 10 kb will be co-repaired on the molecule.

The subject method uses a two vector system where each vector contributes one strand to the double stranded test vector. One contributing vector contains a gene encoding an active, detectable marker. For convenience, this will be referred to as the "A vector". The second contributing vector is substantially complementary to the A vector, except that the marker gene has an inactivating insertion, deletion or substitution loop of at least about 5 nucleotides in length. This vector will be referred to as the "I vector". The A vector and the I vector may be replicated as double stranded DNA, which is then denatured to form single strands, or the vectors may be grown as single stranded entities. The A vector will be replicated under conditions that do not methylate adenine at the GATC recognition site, whereas the I vector will be modified to have methylated adenine at these sites.

One strand from the A vector and one strand from the I vector are annealed to form a heteroduplex, double stranded "A/I" vector. The A/I vector will be methylated on only one strand, *e.g.* the strand that is contributed by the I vector. When the A/I vector is transformed into a suitable bacterial host having an active methyl mismatch repair system (MMR host), the loop between the active and inactive marker gene will not initiate repair. Correction of the marker gene will only take place when there is a mismatch capable of initiating repair elsewhere in the molecule.

The A/I vector is ligated to a "test sequence". The test sequence is a double stranded DNA molecule comprising the sequence of interest, which is being tested for mismatches. A mismatch in the test sequence will initiate repair of the loop in the marker gene in the bacterial host cell. Each strand of the test sequence is contributed by a different source, herein termed X and Y

strands. One or both of the X and Y strands may be ligated to the A and I vectors prior to the previously described annealing step. Alternatively, the X or Y strand(s) is ligated to the double stranded A/I vector. The A/I vector ligated to the test sequence shall be referred to as the test vector.

If the X and Y strands of the test sequence are perfectly complementary, then bacteria transformed with the test vector will not initiate correction of the loop in the marker gene, and will express a mixture of the active and inactive marker. If X and Y are mismatched, then repair is initiated. The marker gene will be "corrected" by co-repair, so that both strands will have the inactive marker sequence. Transformed bacteria will therefore lack active marker. The transformed bacteria are grown on plates, liquid culture, *etc.*, under conditions where expression of the marker can be detected. The presence of transformants that lack the marker indicates a mismatch in the test sequence. These transformants may then be isolated for further use. Figure 1 is a schematic depicting this process.

DNA Vectors

The A and I vectors may be any double stranded or single stranded episomal DNA element that is replicated in the MMR bacterial host, *e.g.* phage, plasmids, bacterial artificial chromosomes (BACs), *etc.* Many vectors are known in the art and are commercially available. The two vectors are substantially complementary if single stranded, and substantially identical if double stranded, except for the previously discussed loop in the marker gene, and optionally, the X or Y sequence of interest. Double stranded vectors must be linearized and denatured prior to formation of the A/I vector. The vectors will contain at least one methylation recognition sequence, generally GATC, more usually multiple recognition sequences will be present..

The A and I vectors have an origin of replication that is active in the MMR host cell. The origin may provide for a high or low copy number of the vector. Optionally, the vectors will include a gene encoding a selectable marker, *e.g.* antibiotic resistance; genes or operons that complement a metabolic defect of the MMR host; resistance to phage infection, *etc.* Phage vectors may include packaging signals, genes encoding phage coat proteins and regulatory genes, *etc.* Desirably, the vector will contain a polylinker having a number of sites for restriction endonucleases to facilitate cloning.

The detectable marker gene may be any gene expressed in the bacterial host that provides a detectable characteristic. Markers of interest include

antibiotic resistance, color change of a substrate, expression of luciferase, *etc.* The use of markers that provide for a color change may be detected by growing the transformed bacteria on medium that allows for the color change, but where the active marker is not required for growth. Transformants expressing the marker are then detectable by visual inspection, spectrophotometry, flow cytometry, *etc.* The use of antibiotic resistance as a detectable marker, *e.g.* expression of β-lactamase, *etc.* will generally require duplicate plates to isolate the mismatched sequence. For example, transformants are grown under non-selective conditions, and a duplicate plate grown under selective conditions. The colonies that cannot grow in the presence of the antibiotic have a mismatched test sequence. A convenient marker is the LacZα gene, which permits the induction of β-galactosidase expression in the presence of isopropyl-β-D-thiogalactoside (see Messing, *supra.*). The β-galactosidase cleaves indolyl-β-D-galactoside to produce a colored product.

The inactivated marker gene on the I vector has an insertion, deletion or substitution "loop" of at least about 5 nt. The minimum size of the loop is required because the loop must not initiate repair by the MMR host. Larger loops, of as much as several hundred bases, may be introduced, but are not necessary for the practice of the invention. The loop inactivates the marker gene by introducing a frameshift, stop codon, *etc.*

In most cases, the I vector will provide the methylated strand. This is done so that during co-repair, the marker gene will be converted to the inactive form. For a number of markers, the active gene is dominant over the inactive. For example, a transformant containing one active antibiotic resistance gene and one inactive gene will be able to grow under selective conditions. Under these same conditions, one can easily distinguish inactive marker from mixed active/inactive. It will be understood by one of skill in the art that this type of a qualitative analysis is merely a convenience, and not essential to the practice of the invention. Methods of quantitative analysis, *e.g.* ELISA, RIA, *etc.*, that can distinguish between the amount of marker produced by one active gene and the amount of marker produced by two active genes (or multiples thereof) may also be used. Such quantitative methods permit either the detection of cells having only active marker from cells having a mixture of active and inactive, or the detection of cells having only inactive marker from cells having a mixture of active and inactive.

The I vector, which is methylated on the adenine of the GATC recognition site, can be replicated in most common laboratory strains of *E. coli*. Other

bacterial hosts that modify DNA at this site may also be used for preparing the I vector DNA. Generally, DNA replicated in non-bacterial cells will require an additional *ex vivo* methylation step, using purified DNA methylases. Substantially all of the GATC sites in the I vector will be methylated. The A vector must be replicated in a host that lacks this DNA modification system. Suitable *E. coli dam-* strains include JM110, described in Janisch-Perron (1985) Gene **33**:103-119. A vectors replicated in non-bacterial host cells, *e.g.* yeast, mammalian cell culture, *etc.* may also be used.

Convenient vectors for preparation of single stranded DNA are derivatives of M13 phage, see Messing (1983) Meth. in Enzym. **101**:20. M13 is a filamentous bacteriophage, and is commonly used in research laboratories. Derivatives of the wild-type phage are known in the art, and commercially available from a number of sources. M13 phage (+) strand DNA can be isolated from phage particles. Double stranded phage DNA is isolated from infected cells, and the (-) strand can be isolated from the double stranded form by various strand separation methods known in the art, *e.g.* columns, gels. Alternatively, the (+) strans may be used in combination with the double stranded form. *E. coli* strains suitable for M13 replication include JM101, JM105, JM107, JM109, *etc.*

The strands of the A and I vector that participate in forming the test vector are substantially complementary. To form the test vector, the A and I vectors are linearized, denatured if necessary, and annealed to each other. Various methods are known for linearizing molecules, *e.g.* digestion with restriction enzymes, *etc.* Methods of denaturing and annealing DNA are well known in the art, and need not be described in detail. The two termini may have blunt ends, or complementary overhanging ends. The annealed, heteroduplex DNA is circularized by a ligation reaction, using any suitable ligase, *e.g.* T4, E. coli, *etc.*, using conventional buffers and conditions. Generally, the quantity of heteroduplex DNA formed will be sufficient to detect in a standard transformation reaction, *e.g.* at least about 0.1 picograms of DNA.

Where double stranded vectors are used, the vectors must be linearized and denatured prior to the annealing step. In addition, it is desirable to remove the homoduplex A and I vectors after annealing and prior to transformation, in order to avoid a high background of transformants. One convenient method of performing this step takes advantage of the differential methylation of the two vectors. Restriction enzymes are known in the art that will cleave homoduplex unmethylated DNA, *e.g.* Mbo I, and homoduplex methylated DNA, *e.g.* Dpn I, but

will not cleave heteroduplex DNA having one methylated and one unmethylated strand. The double stranded A and I vectors are denatured, combined, and reannealed, leaving a mixture of homoduplex DNA (A vector, I vector) and heteroduplex DNA (A/I vector). The mixture is then treated with the methyl specific restriction enzymes. The homoduplex DNA is cleaved, and the heteroduplex is not. The heteroduplex DNA is then used in subsequent steps of the method.

The Test Sequence

The test sequence is a heteroduplex of X and Y, as previously described. X and Y are substantially complementary, and anneal with each other. Generally, the sources of the X and Y strands will be closely related, *e.g.* individuals of a single species, individuals of closely related species, germline and somatic tissue from a single individual, inbred strains of a species, *etc.* The test sequence may be derived from any source, *e.g.* prokaryotic or eukaryotic, plant, mammal, insect, *etc.* The subject method is particularly useful for the analysis of complex genomes, such as those found in higher plants and animals. The test DNA sequence will usually be of at least about 20 nt in length, and usually not more than about 10^4 nt in length. The upper limit on length is determined by the ability of the MMR host to co-repair the strand.

In order to initiate co-repair of the marker gene, there must be at least one "initiating mismatch" in the test sequence. An initiating mismatch is a deletion, insertion or substitution of from one to four contiguous nucleotides. A loop of five or more contiguous nucleotides will not initiate repair. Multiple non-contiguous mismatches may be present in the test sequence. Generally, the test sequence will have at least about 90% identity between the two strands. Initiation of co-repair will proceed as long as one initiating mismatch is present.

Various methods may be used to to generate the X and Y strands. Isolating and amplifying DNA sequences are known in the art. X and Y may be cDNA from a reverse transcriptase reaction, a restriction fragment from a genome, plasmid, YAC, virus, *etc.*; an amplification product from polymerase chain reaction (PCR), *etc.* An important limitation to the use of PCR products is the choice of thermostable polymerase. Polymerases having a 3' to 5' exonuclease activity, *e.g.* proofreading function, are preferred. Useful thermostable polymerases with proofreading capability that are known in the

art include those isolated from *Thermococcus litoralis*, *Pyrococcus furiosis*, and *Thermus thermophilus*. Commercially available *Thermus aquaticus* polymerase has been found to introduce a significant number of errors into the amplified DNA, and will generally be unsuitable for all but very short, *e.g.* less than about 500 nt., sequences.

A number of techniques are known in the art for isolating single strands, or for denaturing double stranded DNA. For example, a reverse transcriptase product may be treated with ribonuclease to leave only the DNA strand. Strand separation gels are known in the art and may be used to separate the two strands of a DNA molecule. PCR may be performed with one primer conjugated to a molecule with a binding partner, such as biotin, haptens, *etc.* The PCR reaction is then denatured, and bound to a solid substrate conjugated to the binding partner, *e.g.* avidin, specific antibody for the hapten, *etc.* The test DNA may be replicated as a single stranded entity, *e.g.* M13 phage, *etc.* The X and/or Y sequence may be restriction fragments, PCR products, or other double stranded DNA molecules, that are denatured according to conventional methods. International application PCT/US93/10722 describes one method for generating heteroduplex DNA suitable for mismatch testing.

There are several different methods that may be used to attach the test sequence DNA to the vector(s). In one method, the double stranded A/I vector is ligated to double stranded X/Y test sequence DNA. In another method, X and Y DNA is ligated into the A and I vectors in a separate cloning step, and the chimeric DNA strands are used to form the A/I heteroduplex molecule. In a third method, test DNA from only one source (X) is cloned into the A or I vector, to form a chimeric molecule. In this case, the heteroduplex A/I vector is gapped, having a single stranded region corresponding to the X sequence. The Y strand is then annealed and ligated into the AX/I vector.

The first method ligates double stranded heteroduplex A/I vector to double stranded heteroduplex X/Y test DNA. The two double stranded DNA molecules are combined. It is convenient to have a short, complementary overhang on the termini of the X/Y, and the A/I molecules, such as those formed by digestion with various restriction endonucleases or by the ligation of specific linkers to the termini, where the vector and the test sequence will anneal to each other. Preferably, a different overhang will be present on each termini of one molecule, so as to prevent self-circularization of the vector. Blunt ends may also be used, in which case it may be desirable to phosphatase treat the vector

ends to reduce self-circularization. The molecules are ligated to form a circular dsDNA, which is then used in subsequent steps.

The X and Y sequences may be separately cloned into the A and I vectors, using conventional recombinant DNA methods (see Sambrook *et al.*, *supra*.). Either strand may go into either vector. The chimeric molecules may then be replicated as previously described, to provide methylated and unmethylated strands. The chimeric molecules are linearized, denatured if necessary, annealed, and ligated as described above to form the A/I vector.

In many cases, it will be desirable to clone only one strand of the test sequence into a vector, and have the other strand of the test sequence be provided separately. Using conventional recombinant DNA techniques, the test sequence (arbitrarily designated X) is cloned into the A or I vector. Either vector may be recipient of the X DNA. For some uses of the method, it may be advantageous to use the A vector as recipient, because the final DNA product, after transformation and methyl mismatch repair, will then be corrected to have the sequence of the Y (methylated) strand, thereby allowing isolation and further growth of the Y DNA. If the vector will be grown as a single stranded entity, then the complementarity of the strands must be selected so that X and Y will be capable of hybridizing.

The chimeric A or I vector, containing X DNA, is linearized and annealed to the complementary vector, to form a heteroduplex A/I vector having a single stranded X region. Y DNA is combined with the heteroduplex vector, and annealed to X. Y may be denatured double stranded DNA, *e.g.* a PCR product, fragment of genomic DNA, *etc.*, or may be single stranded, *e.g.* cDNA, *etc.* The three strands (I, AX and Y) are then circularized and ligated.

Transformation and Detection

The test vector, heteroduplex A/I vector ligated to X/Y test sequence DNA, is transformed into a suitable bacterial host. Most bacterial species have an active methyl mismatch repair system, and can therefore be used as an MMR host. Suitable species include *E. coli* and other gram negative rods, such as *Pseudomonas*, *Erwinia*, *Shigella*, *Salmonella*, *Proteus*, *Klebsiella*, *Enterobacter* and *Yersinia*. Other species of interest include *B. subtilis*, *Streptomyces*, *etc.* The genetics and growth requirements of *E. coli* are well known, and in most cases it will be the preferred host. Transformation techniques are well known, for example see Hanahan (1985) in: DNA Cloning, Vol. 1, ed. D. Glover, IRL Press Ltd., 109.

The transformed bacteria are generally grown under selective conditions, where only those cells able to express a vector encoded selective marker can proliferate. Preferably the test vector will include a selective marker, such as antibiotic resistance, for this purpose. The transformants may be grown in a suitable culture medium, *e.g.* LB broth, SOB broth, 2YT, *etc.*, as a liquid culture, on plates, *etc.* In some cases, the growth medium will also include any substrates required for showing of the detectable marker.

The determination of transformants expressing active and inactive marker is then made. The method of determination will vary with the specific marker used, as previously discussed. In one embodiment, plates of transformants are counted for colonies having a positive or negative color change, such as cleavage of indolyl-β-D-galactoside to produce a blue color, or expression of luciferase. In another embodiment, replica plates are made, and it is determined whether cells from individual colonies are capable of growing in a selective medium. Transformants grown in liquid culture may by stained, for example with antibodies specific for the selectable marker, and analyzed by flow cytometry to determine the number of cells expressing active marker.

Transformants that lack active marker had an initiating mismatch in the test sequence. An increase in the percentage of transformants that lack active marker, compared to a control, prefectly matched test sequence, is indicative of a mismatch. The transformed bacteria that lack active marker are growing the "corrected" test vector, where both strands of vector DNA will have the sequence of the originally methylated strand. The transformed bacteria that express active marker will generally have a mixture of A and I vector. Vector DNA may be prepared from the transformants, and used for further purification and characterization.

Applications of the Method

The subject method is useful for analysis of DNA polymorphisms, and for isolation of variant sequences. A number of applications for the subject method are based on detection of sequence polymorphisms in a single, known DNA sequence. For example, in prenatal diagnosis one might wish to determine whether a mutation in a particular gene, *e.g.* hemoglobin, dystrophin, *etc.*, is found in a fetal DNA sample. Many tumor cells contain a mutation in one or more oncogenes and/or tumor suppressor genes. Determining whether a particular gene is altered in a tumor cell sample is therefore of interest. Determining the occurrence and frequency of sequence

polymorphisms in a population is important in understanding the dynamics of genetic variation and linkage disequilibrium.

To perform this type of analysis, a control (X) copy of the sequence of interest is cloned into the A or I vector, usually A vector. Where a gene is known to be polymorphic, several different vectors, each having a different allelic form, may be used. The Y sequence is obtained from a suitable source of DNA, depending on the type of analysis being performed. The Y sequence may also be cloned into a vector. In a preferred embodiment, however, a heteroduplex is formed of AX and I strands, then combined with single stranded Y DNA, where Y may be a denatured PCR product, cDNA *etc.* X and Y are annealled, and a ligation is performed to produce the test vector.

For human gene mapping, one may set up a panel of A or I vectors having defined regions of a chromosome, for example the BRCA1 gene, or CF gene, where a copy of the gene sequence is cloned into the vector. Due to allelic variation, it may be necessary to compare several sets of control vectors. The length of some genes may necessitate a series of vectors, in order to cover the entire region. The Y sequence DNA is obtained from the individual being tested, using any convenient source of DNA. The Y sequence may be added to the AX/I heteroduplex, or may be cloned into the I vector in a separate reaction. Hybridization of the panel of X sequence vectors with the corresponding Y sequences may be performed in parallel, or in a multiplex reaction. Where a multiplex reaction is performed, the transformed bacteria may be transferred to an ordered array, *e.g.* nitrocellulose, 96 well plate, *etc.*, and analyzed by Southern blot for the presence of any specific sequence. The presence of specific sequences is then correlated with the presence or absence of active marker gene. One can then determine, for large regions of DNA, where an individual sequence varies from a standard, control sequence.

The resulting colonies from the above procedure will be a mixture of active marker expressing, having a DNA sequence identical to the control sequence, and lacking active marker, where there was an initiating mismatch in the test sequence. In order to analyze the results, it may be desirable to determine the frequency of these two populations. This may be accomplished by separating the active and inactive colonies into two different pools. Separation may be accomplished by picking colonies, flow cytometry, column separation based on binding of the marker, immunomagnetic bead separation, *etc.* Vector DNA isolated from these pools is digested with an appropriate restriction endonuclease to release the insert. Gel electrophoresis may then

be used to quantitate the amount of insert DNA in each pool, using the vector band as an internal standard, from which the proportion of variant and identical clones can be determined. Alternatively, the colonies may be transferred to nitrocellulose, and the insert DNA from each of the pools used as a hybridization probe. The ratio of signal intensity from hybridization with the active and inactive pool of inserts can be used to determine the proportion of variant and identical sequences. This allows the simultaneous analysis of sequence variation for many different fragments.

In other applications of the method, one may wish to isolate variants of sequences, particularly genomic sequences. In some cases, the control sequence will be only partially characterized. For example, many genetic diseases or conditions are known only by their phenotype and general map position, *e.g.* a high predisposition to breast cancer, obesity, *etc.* Localization of the gene to a particular map region, or a YAC clone, still leaves hundreds of thousands of bases of DNA containing the potential gene candidate. MRD provides a means of identifying and isolating the variant sequence.

DNA, preferably not more than 2×10^6 bp is isolated from two sources. The DNA may be from a YAC or BAC insert, a restriction fragment from a human chromosome, *etc.* One source of DNA will have the putative variant sequence, and the other will have the control sequence, *e.g.* wild-type. Preferably the two sources will be related, *e.g.* inbred mouse strain, human parent or sibling, *etc.* The transformed cells are useful as a sourced of cloned DNA.

In one method, the two DNA samples are cloned into the I and A vectors, respectively, to provide inserts of not more than about 10^4 nt in length, and usually at least about 10^2 nt in length. The vectors are separately replicated in methylation positive and methylation negative conditions, either as single or double strands. The two vectors are then linearized, denatured if necessary, annealed, ligated, and transformed into an MMR host, as previously described. There will be a large number of transformants that represent perfect matches, and will express active marker gene. The transformants that lack an active marker have a mismatch between the two DNA sources, and are candidates for clones of the variant sequence.

The ability of MRD to islate DNA having a variant sequence can be used in "multiplexing" procedures, where multiple DNA fragments are analyzed in a single reaction. Multiplex reactions may be set up for specific fragments of DNA or regions of a chromosome, *etc.* In multiplex reactions, generally two

cycles of MDR will be performed. The first round of MDR provides a number of bacterial colonies having variant or identical allele(s) from a pool of DNA fragments. The second round of MDR further enriches for the variant sequences.

Regions of up to about 2 megabases of DNA may be compared in multiplex reactions. One or many different fragments may be isolated in a single reaction. Generally the DNA will be fragmented by a suitable method, *e.g.* restriction endonuclease digestion, *etc.*, cloned into the appropriate vectors, and a first round of MRD analysis performed in a single reaction. Colonies having inactive marker after the first round are enriched for variant sequences. DNA isolated from these colonies may be compared to the control sequence, using additional round(s) of MRD to further enrich for variants. The majority of inactive colonies from the second round will carry DNA sequences that differ from the control. Where error prone polymerase was used to generate DNA, the method of "cleaning" described below may be used to enrich for true variants.

An alternative approach to isolating variant sequences is as follows. Two DNA samples, *e.g.* YAC, plasmid, restriction fragment, *etc.*, containing the region of interest are cleaved with a restriction endonuclease into fragments of not more than about 10^4 nt. The two samples are combined, denatured, and allowed to anneal. The X/Y mixture is then anealed and ligated into a heteroduplex A/I vector having compatible ends. The mixture is transformed into an MMR host. Any transformants lacking active marker will represent a mismatch between the two DNA sources.

MRD may be used in conjunction with Taq polymerase to enrich for molecules that are free of PCR-induced errors. Following this "cleaning" protocol, the cloned PCR products is isolated for further analysis. The products of a Taq PCR reaction are cloned into the control and test vectors, and are then hybridized and transformed. The majority of transformants containing Taq PCR-induced errors will present as heteroduplex molecules containing a mismatch and will not produce active marker. In contrast, those PCR products with no PCR-induced errors will contain no mismatches and will produce active marker. These colonies can be isolated, and if desired, undergo a second round of cleansing.

It is contemplated that a kit will be provided for the practice of the subject invention. At a minimum, the kit will contain A and I vectors. The vectors may be single or double stranded. Single stranded vectors may be pre-annealed in

an A/I heteroduplex. Competent host bacteria for growing unmethylated and methylated vector may also be included, as well as an MMR host strain. For analysis of specific DNA sequences, *e.g.* oncogenes, tumor suppressor genes, human β-hemoglobin, cDNA and genomic copies of BRCA1 and BRCA2, a panel covering the human dystrophin gene, *etc.*, a kit may be provided where a chimeric A vector is provided, containing the X (control) sequences. The A and I vector in this case may also be pre-annealed, to form an AX/I heteroduplex. Such a kit may also include specific primers for amplifying the Y sequence DNA, and optionally, thermostable polymerase.

The following examples are offered by way of illustration and not by way of limitation.

EXPERIMENTAL

Two pUC-derived plasmids, the A plasmid (pMF200) and the I plasmid (pMF100), are employed in the MRD procedure. A map of the plasmids is shown in Figure 3. These plasmids are identical except for a five bp insertion into the Lac Zα gene of pMF100. This insertion results in white colonies when bacteria transformed with the I plasmid are grown on LB plates supplemented with indolyl-ß-D-galactoside (Xgal) and isopropyl-ß-D-thiogalactoside (IPTG). In contrast, bacteria transformed with the A plasmid result in blue colonies when grown under these conditions.

The initial step of the MRD procedure consists of cloning one of two DNA fragments to be screened for differences into the A plasmid and cloning of the second DNA fragment into the I plasmid. The A plasmid construct is then transformed into a *dam-* bacterial strain, resulting in a completely unmethylated plasmid while the I plasmid construct is transformed into a *dam*$^+$ bacterial strain, resulting in a fully methylated plasmid. The two plasmids are then linearized, denatured, and reannealed, resulting in two heteroduplex and two homoduplex plasmids. Following digestion with Mbo I and Dpn I, which digest only homoduplexes, the remaining hemimethylated heteroduplexes are circularized, transformed into *E. coli*, and plated onto agar supplemented with Xgal and IPTG.

In the absence of a mismatch between the two test DNA fragments, the five nucleotide loop in the Lac Zα gene, resulting from heteroduplex formation between the I and the A plasmids, is not repaired by the mismatch repair system. Subsequent plasmid replication produces both I and A plasmids in a single colony, leading to a blue color. In contrast, if a mismatch is present in

the heteroduplex DNA, a co-repair event takes place that involves both the mismatch in the DNA as well as the five nucleotide loop in the Lac Zα gene. In this case, the unmethylated Lac Zα gene on the A plasmid is degraded, and replaced by the Lac Zα gene from the methylated strand of the I plasmid, resulting in a white colony. The data show that co-repair of a mismatch and the Lac Zα gene in the MRD system occurs even when the distance between them is greater than 5 kb.

Methods

The MRD vectors. pMF100 and pMF200 are derived from pUC19, with the multiple cloning site displaced from the Lac Zα region. In addition, the MRD vectors contain the Bgl I fragment (2166-472) and most of the multiple cloning site of pBluescript (Stratagene, La Jolla, CA). The cloning sites of the MRD vector do not have sites for the restriction endonucleases XbaI, SpeI, BamHI, SmaI and ApaI. The EcoRI site is not unique. pUC19 multiple cloning istes, nucleotides 400-454, were replaced using 70 nucleotide long oligonucleotides with a sequence containing four GATC sites. In addition, the site replacing the pUC19 multiple cloning sites in pMF200 has a 5 bp insertion as compared to pMF100, creating a non-functional Lac Zα in pMF200. The label "loop" in Figure 3 indicates this difference.

FOrmation of heteroduplex DNA. DNA from the unmethylated and methylated plasmids are linearized, denatured, and reannealled. The resulting molecules are fully unmethylated A plasmid homoduplexes, fully methylated I plasmid homoduplexes, and hemimethylated heteroduplexes. The mixture is digested with MboI, which digests fully unmethylated DNA, and DpnI, which digests fully methylated DNA. Only the heteroduplex, hemimethylated DNA is left.

Example 1

As an initial test of the sensitivity and specificity of the MRD system, a single nucleotide mismatch was detected in a 550 base pair DNA fragment derived from the promoter of the mouse beta globin gene (Myers *et al.* (1985) Science **229**:242). MRD was used to compare this DNA fragment, which contains a T at position -49 (relative to the functional transcription start site of the gene) with a second DNA fragment identical in sequence except for at C position -49. The mismatch was located about 700 base pairs from the five nucleotide Lac Zα loop in the vector. Comparison of the two DNA molecules by using MRD resulted in 90% white colonies. In contrast, comparison of the

same two DNA molecules with no mismatch (-49T/-49T), resulted in only 7% white colonies. The data is shown in Table 1.

Table 1
Detection of Known Point Mutations using MRD

Sequence Variation*	Fragment Size^	Distance from Loop^	% White (Inactive) Colonies@
none[1]	0.55	N/A	7
G→C[1]	0.55	0.7	89
A→T[1]	0.55	0.7	84
G→T[1]	0.55	0.7	82
A→C[1]	0.55	0.7	82
C→T[1]	0.55	0.7	90
none[2]	2.0	N/A	8
A→C[2]	2.0	0.4	35
none[3]	2.2	N/A	10
C→T[3]	2.2	2.3	83
G→A[3]	2.2	2.1	86
C→T[3]	2.2	1.6	81
T→C[3]	2.2	1.8	80

* A→T variation means that at the only position of ariation between the two fragments compared, the dam- grown variant has an A and the dam+ grown variant has a T at the same position on the same strand. Therefore, mismatches produced in such an experiment are A/A and T/T.
^ in kilobases.
@ At least 250 colonies were counted to determine the percentage.
1. Experiment using a fragment of the mouse beta globin gene.
2 Experiment using a fragment of the human agouti gene.
3. Experiment using fragment of human cystathionine beta synthase gene, at positions 341, 502, 992, and 833, respectively.

Comparison of all possible single nucleotides mismatches at position -49 using MRD revealed proportions of white colonies ranging from 80% to 90%. These results demonstrate that MRD can detect all of the different DNA variations possible at this position with high efficiency.

The MRD system was used to detect a total of five additional single nucleotide mismatches in two different DNA fragments, shown in Table 1. Four of these mismatches are at different nucleotide positions in the human cystathionine beta synthase gene (Kruger and Cox (1995) Human Molecular Genetics **4**:1155). The remaining one mismatch represent single nucleotide changes in the human agouti gene (Wilson *et al.* (1995) Human Molecular Genetics **4**:223). In each case, a single nucleotide mismatch was detected.

A mismatch was detected even when it was as far as 2.3 kb from the Lac Zα loop. Since the proportion of white colonies was greater than 50%, co-repair of the mismatch and the loop on the unmethylated strand occurred irrespective of which side of the mismatch was relative to the loop.

To determine whether the efficiency of mismatch detection would remain high if the distance between a mismatch and the vector loop was even larger, the following experiment was performed. A 9 kb test DNA fragment derived from lambda bacteriophage was cloned into the MRD plasmid system and compared with the same test DNA containing a two base pair insertion located 5 kb from one end of the fragment. Addition of the two base pair mismatch resulted in 70% white colonies, as compared to 10% white colonies in the absence of the mismatch. These results indicate that MRD can detect a mismatch in 10 kb of DNA.

Example 2

MRD was used to detect unknown mutations in genomic DNA fragments generated by the polymerase chain reaction (PCR). PCR is a practical method for obtaining a particular genomic DNA fragment of interest from many different individuals. Recent advances in PCR technology makes it possible to isolate DNA products greater than 10 kb in length (Barnes (1994) P.N.A.S. **91**:2216; Cheng *et al.* (1994) P.N.A.S. **91**:5695). However, the introduction of errors during the PCR reaction severely limits the use of individual cloned PCR products. In an effort to overcome this limitation, an MRD protocol was developed to enrich for molecules that are free of PCR-induced errors. Following this "cleaning" protocol, the cloned PCR products can be compared for DNA sequence differences by using the MRD procedure described above.

The basic principle underlying the MRD cleaning protocol is the fact that any single PCR-induced mutation will make up a very small fraction of all the molecules generated by PCR. As a result, when the products of a PCR reaction are cloned into the A "blue" and the I "white" MRD vectors and assayed

as described above, the majority of products containing PCR-induced errors will present as heteroduplex molecules containing a mismatch and will produce white colonies. In contrast, those PCR products with no PCR-induced errors will contain no mismatches and will result in blue colonies. Given that not all mismatches are repaired with 100% efficiency, some blue colonies can be expected to contain PCR-induced errors following the first round of enrichment. However, if blue colonies are isolated and used in a second round of MRD cleaning, those molecules containing PCR-induced errors can be reduced even further. Since each blue colony contains both a blue MRD plasmid and a white MRD plasmid, the second round of MRD cleaning is carried out as follows. Plasmid DNA isolated from blue colonies following the firs round of cleaning is used to transform both dam- and a dam+ bacterial strains. Although both blue and white colonies resulted from each transformation, only the blue colonies are isolated from the dam- transformation, and only the white colonies are isolated from the dam+ transformation. Plasmid DNA is prepared from such colonies and heteroduplexes are isolated as described above. Blue colonies arising from transformation with these heteroduplexes are further enriched for the products free of PCR-induced error. In an experiment in which 75% of molecules contain one or more PCR-induced errors following PCR, assuming 95% efficiency of mismatch repair and 10% frequency of white colonies in the absence of a mismatch, the expectation would be 10% blue colonies following one round of MRD enrichment, with 66% of the molecules in such colonies free of PCR-induced errors. If the plasmid DNA from the blue colonies were used for a second round of MRD enrichment, the expectation would be 41% blue colonies, with 96% of the molecule in such colonies free of PCR-induced errors.

As a test of the practicality as well as the efficiency of the MRD cleaning protocol, a 2 kb human chromosome 21-specific PCR product was isolated from each of the two chromosome 21 homologues of a single individual. The two chromosome 21 homologues were separated from each other in independent hamster-human somatic cell hybrid clones. Genomic DNA isolated from these somatic cell hybrid clones was the source of PCR products. When the PCR products derived from each homologue were compared using MRD as described above, approximately 10% blue colonies were observed in each case.

Following two rounds of MRD cleaning, the proportion of blue colonies as 60-80%, data shown in Table 2. In contrast, when these "cleaned" PCR products derived from the two homologues were compared with each other by using MRD, approximately 90% of the resulting colonies were white, indicating the presence of at least one single base difference in the 2 kb PCR products derived from the two different chromosome 21 homologues. The DNA sequence variation in the PCR products was independently verified by restriction enzyme digestion. These results demonstrate that MRD can be used to enrich for PCR products that are largely free of PCR-induced errors, and that such products can be used in conjunction with MRD to detect human DNA sequence variation.

Table 2.
Percentage of Inactive Colonies in Different Comparison with Plasmids containing 2 kb PCR Products from two Somatic Cell Hybrids

Variants Compared*	Percentage of Inactive Colonies#
1/2	>90
2/2	>90
A1/A1	70
A2/A2	64
AA1/AA1	38
AA2/AA2	21
AA1/AA2	>90
AA2/AA1	>90

* 1 and 2 represent products from the two hybrids. 1/1 represents comparison of A vector grown in a dam- strain and containing the PCR product from hybrid 1 to I vector grown in a dam+ strain and containing the PCR product from hybrid 1. A1/A1 represents the comparison of A vector grown indam- host, obtained from the active colonies of comparison 1/1, to I dam+ grown vectors obtained from the same source. AA1/AA1 represents the comparison of A dam- grown vectors obtained from the active colonies of the comparison A1/A1 to I dam+ grown vectors from the same source. Finally, AA1/AA2 represents the comparison of A dam- grown plasmids obtained from the active colonies of the comparison A1/A1 to I dam+ grown vectors obtained from the active colonies of the comparison A2/A2.

It is evident from the above results that the subject invention provides for an efficient, simple method of detecting mismatches between two DNA

sequences. The method provides a means of simply detecting the presence of a mismatch, or can be used to isolate copies of both matched and mismatched DNA. MRD is useful to determining somatic changes in gene sequence, identifying germline mutations for prenatal or other genetic screening, for human gene mapping, and for cloning mutations. A major advantage of MRD is the potential of this system to analyze many fragments simultaneously in a single experiment, allowing the detection of mutations in a region representing hundreds of kilobases of DNA, or for genotyping many loci simultaneously. MRD provides a powerful technique for the detection of unknown mutations, the detection of DNA variation in large genomic regions, and high-throughput genotyping.

All publications and patent applications cited in this specification are herein incorporated by reference as if each individual publication or patent application were specifically and individually indicated to be incorporated by reference.

Although the foregoing invention has been described in some detail by way of illustration and example for purposes of clarity of understanding, it will be readily apparent to those of ordinary skill in the art in light of the teachings of this invention that certain changes and modifications may be made thereto without departing from the spirit or scope of the appended claims.

WHAT IS CLAIMED IS:

1. A method of detecting a mismatch between two substantially complementary DNA sequences of interest, the method comprising:

annealing a first strand comprising a gene encoding a detectable marker and an origin of replication active in a bacterial host cell, wherein said first strand is characterized by the absence of methyl adenine; and a substantially complementary second strand, wherein said gene encoding said detectable marker further comprises an inactivating insertion, deletion or substitution of at least about 5 nt, and characterized by the presence of methyl adenine at GATC sites;

ligating said first strand to a first DNA sequence of interest of from about 20 to 10^4 nucleotides in length;

ligating said substantially complementary second strand to a second DNA sequence of interest substantially complementary to said first DNA sequence, and suspected of having at least one mismatch of from 1 to 4 contiguous nucleotides in length;

circularizing said ligated first and second strands to provide a circular double stranded DNA molecule;

transforming a bacterial host having an active methyl mismatch repair system with said circular double stranded DNA molecule;

detecting the presence of bacterial transformants not expressing said detectable marker;

wherein the presence of transformants not expressing said detectable marker is indicative of a mismatch between said first DNA sequence of interest and said second DNA sequence of interest.

2. A method according to Claim 1, further comprising isolating and growing said bacterial transformants.

3. A method according to Claim 1, wherein said first strand and said substantially complementary second strand further comprise a selectable marker, and a polylinker having multiple sites for restriction endonucleases.

4. A method according to Claim 1, wherein said first DNA sequence of interest is a polymerase chain reaction product.

5. A method according to Claim 1, wherein said first DNA sequence of interest is a cDNA.

6. A method according to Claim 4, wherein said first DNA sequence of interest is a restriction fragment.

7. A method according to Claim 1, wherein said second DNA sequence of interest is a polymerase chain reaction product.

8. A method according to Claim 1, wherein said second DNA sequence of interest is a cDNA.

9. A method according to Claim 1, wherein said second DNA sequence of interest is a restriction fragment.

10. A method according to Claim 1, wherein said ligating said first strand to a first DNA sequence of interest is performed prior to said annealing step.

11. A method according to Claim 1, wherein said ligating said first strand to a first DNA sequence of interest is performed after said annealing step.

12. A method according to Claim 1, wherein said ligating said substantially complementary second strand to a second DNA sequence of interest substantially complementary to said first DNA sequence is performed prior to said annealing step.

13. A method according to Claim 1, wherein said ligating said substantially complementary second strand to a second DNA sequence of interest substantially complementary to said first DNA sequence is performed after said annealing step.

14. A double stranded DNA composition comprising:

a first strand comprising a gene encoding a detectable marker and an origin of replication active in a bacterial host cell, wherein said first strand is characterized by the absence of methyl adenine; and

a substantially complementary second strand, wherein said gene encoding said detectable marker further comprises an inactivating insertion, deletion or substitution of at least about 5 nt, and characterized by the presence of methyl adenine at GATC sites.

15. A DNA composition according to Claim 14, wherein said first strand and said substantially complementary second strand further comprise a selectable marker, and a polylinker having multiple sites for restriction endonucleases.

16. A DNA composition according to Claim 14, wherein said first strand further comprises a first DNA sequence of from about 20 to 10^4 nt in length.

17. A DNA composition according to Claim 15, wherein said second strand further comprises a second DNA sequence substantially complementary to said first DNA sequence, and suspected of having at least one mismatch of from 1 to 4 contiguous nucleotides in length.

18. A kit for identifying the presence of a mismatch between two substantially complementary DNA sequences, the kit comprising:

a first DNA vector comprising a gene encoding a detectable marker and an origin of replication active in a bacterial host cell, and

a second DNA vector substantially identical to said first DNA vector, wherein said gene encoding said detectable marker further comprises an inactivating insertion, deletion or substitution of at least about 5 nt.

19. A kit according to Claim 18, wherein said first and said second DNA vectors further comprise a selectable marker, and a polylinker having multiple sites for restriction endonucleases.

FIGURE 1

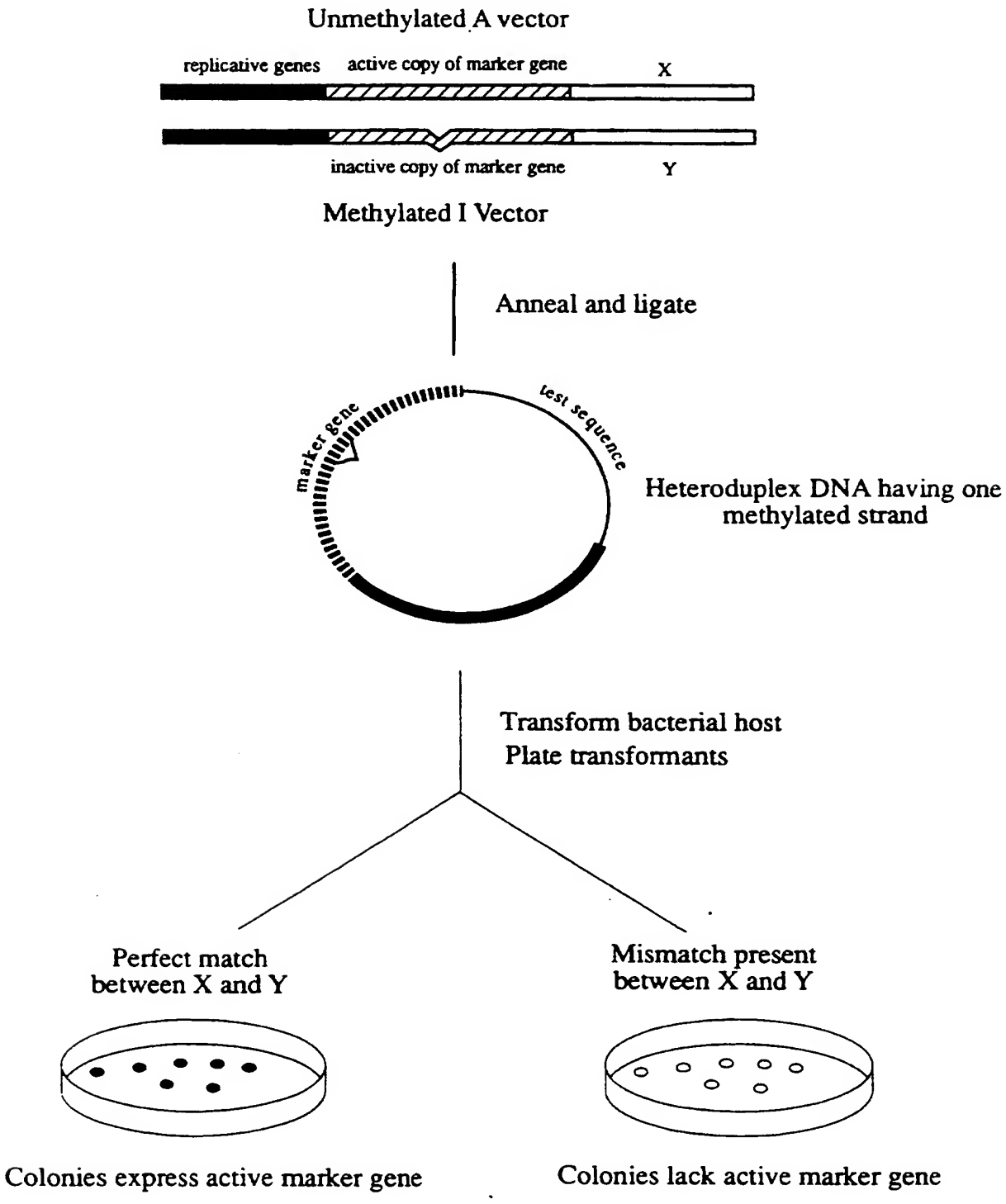


Figure 2

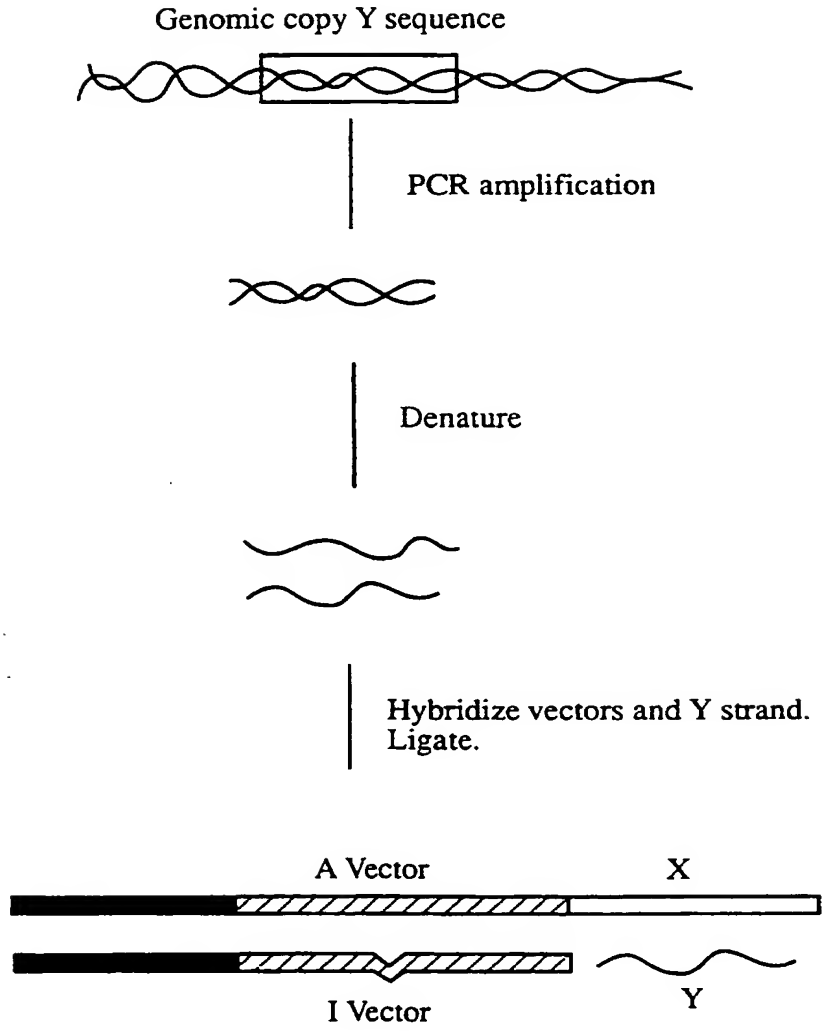

Genomic copy Y sequence
PCR amplification
Denature
Hybridize vectors and Y strand.
Ligate.
A Vector
X
Y
I Vector

Figure 3

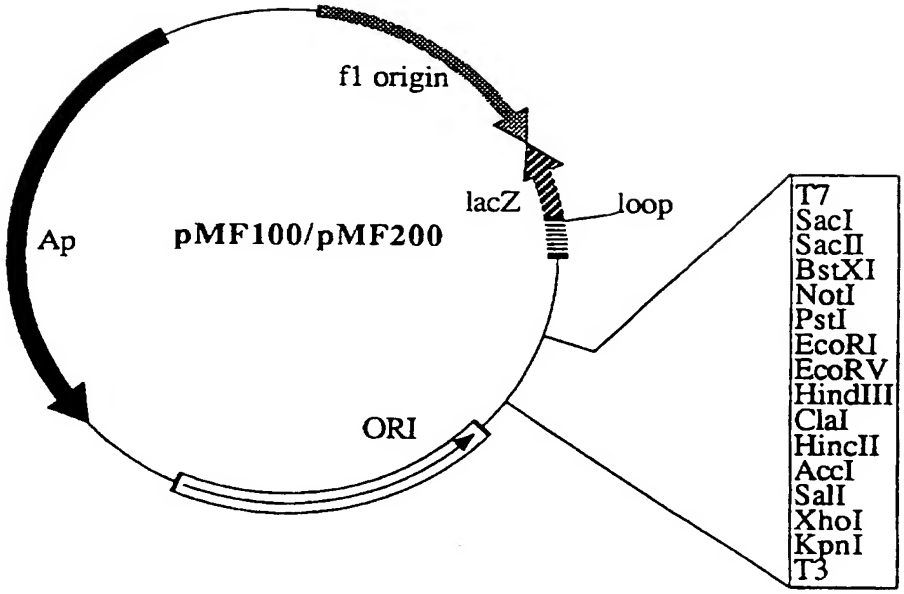

f1 origin
lacZ
loop
Ap
pMF100/pMF200
ORI
T7
SacI
SacII
BstXI
NotI
PstI
EcoRI
EcoRV
HindIII
ClaI
HincII
AccI
SalI
XhoI
KpnI
T3

Figure 4

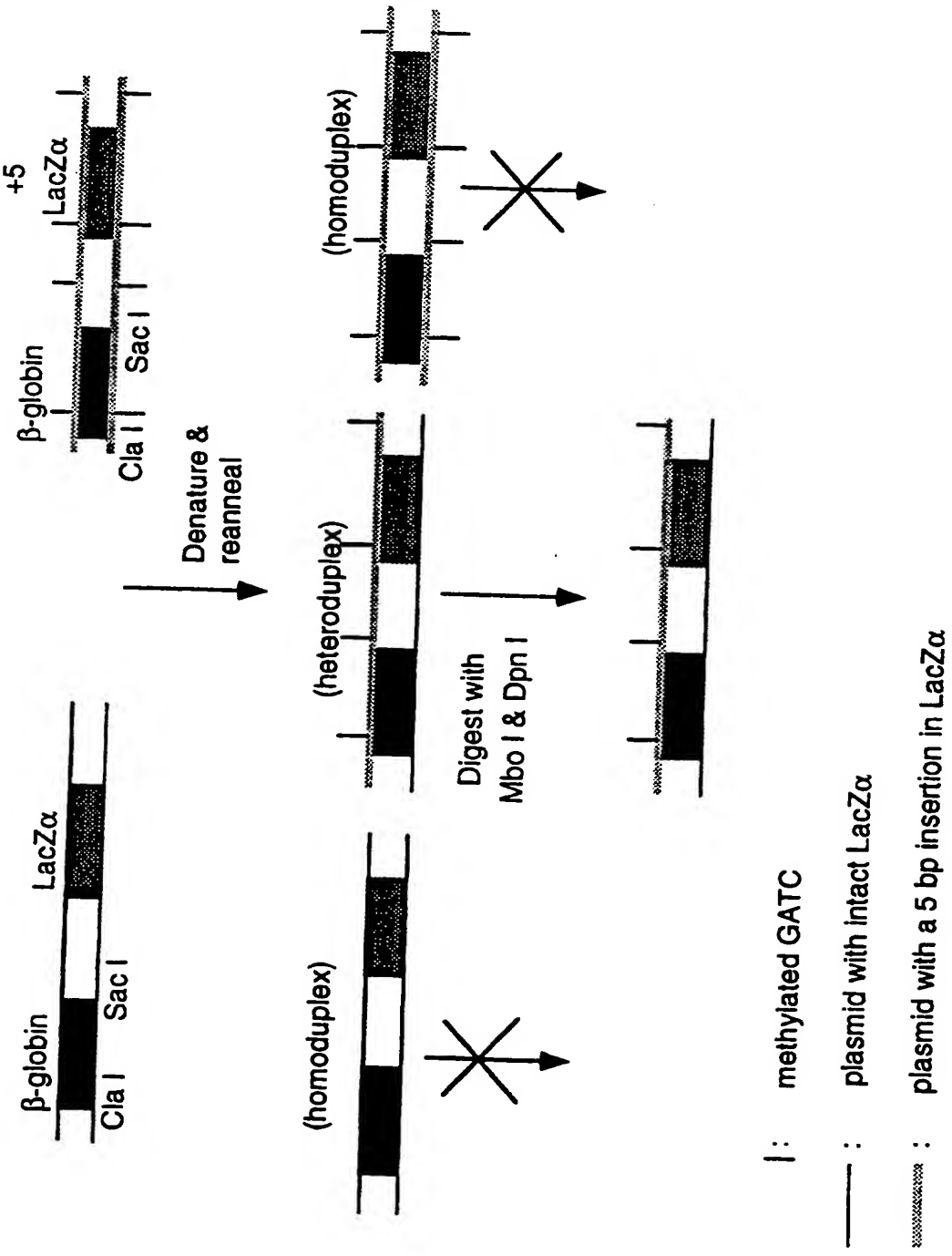
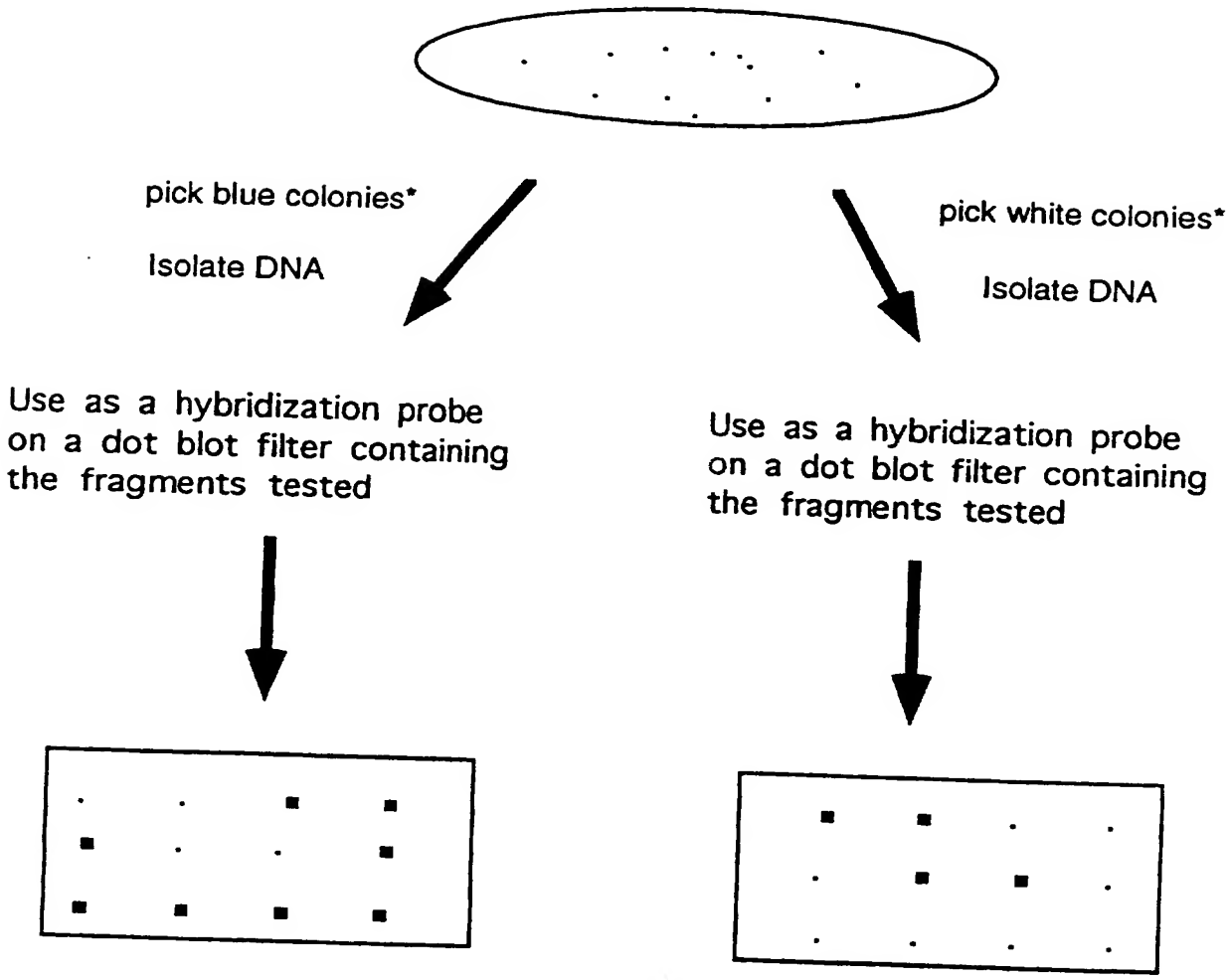

+5
LacZα
β-globin
Cla I
Sac I
Denature & reanneal
(homoduplex)
(heteroduplex)
Digest with Mbo I & Dpn I
(homoduplex)
LacZα
β-globin
Cla I
Sac I
: methylated GATC
: plasmid with intact LacZα
: plasmid with a 5 bp insertion in LacZα

Figure 5



comparing the hybridization signal on both filters, one can determine which fragments are variant

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/14655

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :C12Q 1/68; C12N 1/20, 15/00; C07H 21/04
US CL :435/6, 252.3, 320.1; 536/23.1
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6, 172.1, 252.3, 320.1; 536/23.1, 23.4, 23.7; 436/94; 935/77, 78, 79, 80, 82

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Extra Sheet.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5,354,670 A (NICKOLOFF ET AL.) 11 October 1994, see entire document.	1-19

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
21 NOVEMBER 1996

Date of mailing of the international search report
26 DEC 1996

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. (703) 305-3230

Authorized officer
BRADLEY L. SISSON
Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/14655

B. FIELDS SEARCHED
Electronic data bases consulted (Name of data base and where practicable terms used):

APS

SEARCH TERMS: mismatch, methyl adenine, dna, transformant!, marker gene, pcr, polymerase chain reaction, insertion, deletion, substitution, inactivation, vector, host cell, bacteria, transformant